

This PDF is a selection from an out-of-print volume from the National Bureau of Economic Research

Volume Title: The Role of the Computer in Economic and Social Research in Latin America

Volume Author/Editor: Nancy D. Ruggles, ed.

Volume Publisher: NBER

Volume ISBN: 0-87014-260-7

Volume URL: <http://www.nber.org/books/rugg75-1>

Publication Date: 1975

Chapter Title: Data Preparation for Latin American Comparisons of Consumption

Chapter Author: Howard Howe, Roberto Villaveces

Chapter URL: <http://www.nber.org/chapters/c3777>

Chapter pages in book: (p. 269 - 290)

DATA PREPARATION FOR LATIN AMERICAN COMPARISONS OF CONSUMPTION*

HOWARD HOWE AND ROBERTO VILLAVECES
The Brookings Institution

INTRODUCTION

Effective empirical research depends on reliable data. In sample-survey work, considerable attention has been devoted to the prevention, detection, and correction of errors introduced in the survey process itself. However, there has been little explicit discussion of the problems in accurately transmitting information from the field interviews to forms suitable for statistical treatment. We attempt to focus attention on this question through discussion of data preparation for a major study of Latin American income and expenditure patterns.

The ECIEL¹ household income and expenditure study involves a large-scale body of sample-survey data. The collaborative nature of the study led to decentralized decision making in project design and execution. Computing played a fundamental role not only in the conventional processing of the data, but also in compensating for the organizational complications inherent in joint comparative work.

We establish a framework of technical criteria for the organization of a large-scale study. Project feasibility necessitated relaxation of the technical efficiency criteria at various stages in the project. Foremost among the organizational complications was the awkward position of a U.S. institution coordinating independent Latin American institutes in empirical research. Individual methodological preferences and opportunities for cost shifting contributed to inefficiencies. The large size of the project magnified the effect of uncertainty.

We discuss in depth two major aspects of the data preparation: the conversion of national data to a common code for international comparability, and the data cleaning. This preparatory work entailed considerable computing effort. The data base was fully standardized to reduce the marginal cost of future comparative studies. Decentralization of the fieldwork necessitated greater attention to data scrutiny than would be the case with purely national studies.

We conclude with recommendations for efficiency in comparative survey studies and suggestions for improved reporting on current large-scale computing work.

* The ECIEL program is supported by funds from the Ford Foundation, the Interamerican Development Bank, and the Brookings Institution, in addition to local resources provided by the participating institutions.

We wish to thank Ximena Cheetham, Robert Ferber, Joseph Grunwald, Marcia Mason, Arturo Meyer, Philip Musgrove, and Robert Summers for their helpful comments and criticisms of earlier drafts of the paper.

¹ ECIEL is the Spanish acronym for Joint Studies on Latin American Economic Integration. The program is carried out by twenty independent research institutions in Latin America. The thirteen institutes participating in the income and expenditure study are listed in the Appendix. The Brookings Institution acts as coordinator.

ORGANIZATION OF THE STUDY

The data-processing system depended not only upon the research goals of the study, but also upon its organization. Understanding of the institutional context, in turn, requires some familiarity with the ECIEL program.

In 1963, several major economic-research institutions in Latin America joined in a common research program. They committed themselves to collaborate in comparative economic studies on Latin American integration and development, under the coordination of staff members of the Brookings Institution. The program's major objective was the preparation of professionally competent and relevant empirical studies. The strengthening of the economics profession in Latin America has been an important by-product of this cooperative effort.

Since 1963, additional research institutions have joined ECIEL. Twenty institutions from twelve Latin American countries currently participate in the program. The program is coordinated through twice-yearly seminars, attended by the principal researchers concerned with the ECIEL projects. The Coordinator provides methodological, technical, and administrative support. At the seminars, participants select and design studies, develop methodology and procedures, and resolve research and coordination problems. The seminar site rotates among the participating institutions in Latin America. Periodic visits by Brookings coordinating staff, consultants, and technical specialists from the institutes supplement the seminars. The income and expenditure study is one of four ECIEL studies currently under way.

As the foregoing discussion implies, the organization of the income and expenditure study was decentralized. The institutes participated in the research design from the inception of the study. Sample design and questionnaire design were carried out with the assistance of consulting specialists. The institutes assumed responsibility for the control and, in most cases, the implementation, of the field-survey work and data punching. The actual data processing was centralized to insure international comparability of the data. The Brookings Institution provided central processing facilities.

As a framework for discussing the tradeoffs in the data preparation, we cite three organizational criteria for efficient data processing: (1) uniform questionnaire; (2) centralized data processing; and (3) close linkage between the stages of the study. At all stages of the income and expenditure study, technical efficiency had to be balanced against feasibility within the context of the ECIEL program. Each of the criteria had to be relaxed in some measure.

Uniform Questionnaire

A uniform questionnaire facilitates efficient processing. International comparison is best attained by uniform definition of variables and common means of data collection. In the income and expenditure survey, however, country questionnaires differed substantially.

The Coordination provided an outline questionnaire core. Development of the actual questionnaire was carried out by the institutes. Some institutes had built up vested interests in particular approaches. The differences over collection focused on the diary method as opposed to the recall method. Each technique had committed

proponents. In addition, the institutes were much more conscious of their unique national characteristics than of the wide range of regional similarities. They felt that the presence, or absence, of certain consumer items in each country necessitated country-specific expenditure categories. In this situation, adoption of a uniform questionnaire could have been accomplished in one of two ways, neither of which is attractive: imposition of a questionnaire by Brookings staff, or the direct supervision of questionnaire preparation in each country. It would have been politically unacceptable to impose a U.S.-designed questionnaire for large-scale application by independent research institutions in Latin America. Sensitivity to the second alternative would also have been high, and, in addition, the on-site supervision would have been prohibitively expensive.

To have insisted upon a common questionnaire at this stage of the study would, in all likelihood, have resulted in several countries dropping out of the study. Hence, in the interest of comprehensive coverage, the institutes were left to interpret individually the core outline and *ex post* means to attain comparability were adopted. The national data were transformed to a uniform international code. Essentially, commonality was attained by treating the data at a higher level of aggregation than in the questionnaire.

The increased cost due to the conversion step was relatively independent of the extent of the questionnaire differences. The major cost stemmed from simply allowing for the possibility of differences. Specification of the conversions cost more than their execution. Approximately four man-years of effort on the part of the coordinating staff were necessary to specify the variable-by-variable correspondences between each of the eleven questionnaire codes and the uniform international code. Granted the necessity of permitting different questionnaires, it is nevertheless important to recognize the consequent additional costs in time and resources.

Centralized Data Processing

Large economies of scale can be attained with centralized processing. Uniform treatment of the data, desirable in a comparative study and essential when different questionnaires were employed, is assured with centralized processing. Centralized processing also becomes more important if skill levels differ.

The institutes' capabilities for undertaking a study of this magnitude varied widely. Some had extensive experience in sample surveys, while for a few of the younger institutes, the ECIEL survey was a first experience. Capacities for handling a large data base were also disparate. While some institutes lacked facilities even for keypunching, others enjoyed a full complement of hardware, software, and experienced personnel. Technical support at the data-processing stage was necessary to strengthen the contribution of the less-experienced institutes.

The case for central processing is not all positive. In view of the training objectives of the ECIEL program, local processing would be desirable. The best way to advance Latin American research capabilities would have been to carry out all processing at the local level, with the technical guidance of consulting specialists. However, considering the standardization required by the comparative nature of the study, and the prohibitive cost of replicating the information-processing system eleven times, centralized processing was established as a technical criterion.

To a great extent, centralized processing was attained. After the study was under way, only three institutes felt that they could not export raw information for processing. In all three cases, the surveys had been carried out by government agencies in collaboration with the ECIEL institutes. Superficial similarities between the ECIEL survey and market surveys placed the institutes in a position vulnerable to criticism, albeit unfounded, for serving foreign commercial interests. Sensitivity to the potential for such criticism almost certainly played a part in the decision to process within the country.

Significantly, the countries carrying out the processing locally were among those best equipped with computing facilities. The processing system had to be adapted for use in those countries. Were the data-processing system fully automated, the only problem would have been modification of the programs to function at another installation. However, the processing system employed in the study involved considerable human-machine interaction. It was necessary to train researchers and assistants at the local level to use the data-processing system. It is doubtful that local personnel could have attained the expertise of a central processing staff benefiting from experience in processing other countries' data. The costs of local personnel, modification of the programs to run on a different system, and visits by programming specialists to set up the system, were considerable.

For the processing at Brookings, best results were obtained when institutes sent their programmer or researcher with the data. This provided the optimal combination of firsthand knowledge of the data, experience with the data-processing system, and use of the computer system for which the programs were written. The training objectives of the ECIEL program were also advanced by such arrangements.

Close Linkage Between Stages of the Study

Three main stages are relevant here: field survey, data processing, and analysis. Close linkage provides the intimate contact necessary to take into account the great externalities between stages. Indeed, in self-contained studies, a single organization normally has responsibility for all stages of the study. Often, the same group oversees each stage. The nature of the ECIEL study did not permit the overlap of personnel to attain close linkage. It was the task of the coordinating staff to orchestrate the various stages.

The impracticality of direct supervision of the fieldwork has already been discussed. Since the processing was centralized, it came under direct supervision of the Coordination. (Where processing was carried out locally, the system and programs were provided by the Coordination.) Appreciable coordinating influence could be exerted at the analysis stage. While the institutes were responsible for the analysis of their processed information, the Coordination provided technical assistance and consultation with recognized experts to insure high academic standards. Thus, the field survey was the only area where close communication was not attained.

Extensive data checking was undertaken to reinforce the relatively loose linkage between the field-survey and data-processing stages. Seven kinds of checking, undertaken in two stages, accomplished extremely close scrutiny of the data:

Stage I. Mechanical consistency: (1) sequence check; (2) nonnumeric characters; and (3) embedded blanks.

Stage II. Substantive consistency: (4) valid codes; (5) valid quantitative values; (6) logical and arithmetic consistency; and (7) extreme value test.

This checking was a relatively costly procedure; it constituted a major portion of the data-preparation work.

THE DATA BASE

The ECIEL household income and expenditure study constitutes a benchmark for inter-American comparisons in the consumption sector. The study is based on

TABLE 1
CHARACTERISTICS OF THE SURVEYS

Country and City	Total Population (millions)	Number of Observations	Number of Intervals	Panel	Date of Survey
Argentina	23.6				
Buenos Aires	7.7	1,398	4	Yes	4/29/69; 7/15/70
Bolivia	4.7	1,295			
La Paz	0.5	695	4	No	12/2/67; 5/30/69
Cochabamba	0.3	600	4	No	12/2/67; 5/30/69
Brazil	88.2	2,428			
Rio de Janeiro	4.2	1,006	4	Yes	5/20/67; 9/20/68
Porto Alegre	0.9	706	4	Yes	5/12/67; 5/30/68
Recife	1.1	716	4	Yes	7/26/67; 5/10/68
Chile	9.4				
Santiago	2.4	3,379	4	Yes	9/15/68; 9/30/69
Colombia	19.8	2,949			
Bogotá	2.0	798	4	Yes	2/10/67; 4/30/68
Barranquilla	0.8	727	4	Yes	2/10/67; 4/30/68
Cali	0.8	634	4	Yes	2/10/67; 4/30/68
Medellin	1.0	790	4	Yes	2/10/67; 4/30/68
Ecuador	5.7	1,994			
Quito	0.5	934	4	Yes	5/26/67; 11/14/68
Guayaquil	0.7	1,060	4	Yes	6/17/67; 12/14/68
Mexico	47.3	5,070	1	No	4/68
Mexico, D.F.					
Guadalajara					
Monterrey					
Paraguay	2.2				
Asunción	0.3	566	2	Yes	9/70
Peru	12.8				
Lima	2.4	1,357	4	No	2/15/68; 2/15/69
Uruguay	2.8				
Montevideo	1.2	1,135	4	Yes	8/20/67; 8/17/68
Venezuela	9.7	2,123			
Caracas	2.1	948	1	No	10/15/66; 11/15/66
Maracaibo	0.6	1,175	4	Yes	6/19/67; 3/10/68
Total number of observations		23,694			

large-sample cross-sectional surveys in twenty-one urban centers in all eleven LAFTA² countries. In most countries the surveys covered four quarterly intervals; the surveys of all the countries were carried out between 1966 and 1971. Approximately 24,000 interviews on 19,000 independent households (the difference represents repeat interviews of panel households) were obtained. Table 1 provides a detailed breakdown of the sample sizes by city and the survey dates.

Scope and Objectives of the Study

Knowledge of consumption patterns in Latin America is quite sparse. For example, the national accounts of most Latin American countries estimate private consumption as a residual. This approach does not constitute reliable measurement. Neither does it provide an indication of the composition of private consumption. The ECIEL study seeks to establish base-point national consumption studies. Specific comparative research topics include:

- (1) allocation of family budgets by type of expenditure;
- (2) price elasticities of demand for particular goods;
- (3) income and expenditure elasticities;
- (4) income distribution by level and type of income;
- (5) estimation of cross-sectional consumption functions;
- (6) international comparison of the determinants of consumption expenditures and investigation of the relevance of existing differences for the economic integration of the region;
- (7) expenditure weights for price indexes; and
- (8) estimates of household saving.

Many of these topics are relevant for intracountry comparisons by city and socioeconomic class, as well as for international comparisons.

Common Code

A typical questionnaire contains between 800 and 1,000 variables. The international code to which all country data are transformed contains 1,283 variables. A great many expenditure categories of the questionnaires are collapsed into aggregated variables in the common code. The common code also establishes a number of subtotals and dummy variables in addition to the original information.

Socio-demographic Data

For each member of the consumption unit, the following information is available: sex, age, marital status, education, occupation and occupational status, and relationship to the head of the unit. For income earners, sector of employment and income are also included. Limited migration information (number of years in city and previous place of residence) exists for the head of the unit and spouse.

Considerable information is available for the dwelling: type of dwelling, number of rooms, utility services, ownership status, rent (actual or imputed), and mortgage payments.

² The Latin American Free Trade Association (LAFTA) comprises Argentina, Bolivia, Brazil, Chile, Colombia, Ecuador, Mexico, Paraguay, Peru, Uruguay, and Venezuela.

TABLE 2
CLASSIFICATION OF EXPENDITURE DATA AND INCOME DATA

Expenditure Data: Major Groups and Subgroups	
1. Food and beverages:	
1.01	Dairy products
1.02	Cereals
1.03	Meat and poultry
1.04	Seafood
1.05	Vegetables
1.06	Fruits
1.07	Fats and oils
1.08	Sweets
1.09	Tea, coffee, and hot beverages
1.10	Alcoholic beverages
1.11	Other beverages
1.12	Other foods
1.13	Food and beverages outside the home
2. Housing:	
2.01	Own residence
2.02	Other residences
2.03	Maintenance
3. Household equipment and maintenance:	
3.01	Durable goods
3.02	Non-durable goods
3.03	Services
4. Clothing:	
4.01	Men's
4.02	Women's
4.03	Children's
4.04	Other clothing
5. Medical care	
6. Education	
7. Recreation and cultural activities:	
7.01	Entertainment
7.02	Reading material
8. Transportation and communication:	
8.01	Own transportation: purchase of vehicle (net outlay)
8.02	Own transportation: current expenses
8.03	Public transportation
8.04	Telephone and other communication
9. Other consumption expenditures:	
9.01	Tobacco
9.02	Personal care
9.03	Entertainment
10. Taxes	
11. Insurance:	
11.01	Social security
11.02	Other insurance
12. Transfers, gifts	
13. Miscellaneous non-consumption expenditures	
14. Total expenditure	$(1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 + 9 + 10 + 11 + 12 + 13)$
Income Data: Major Groups	
1. Wage income	
2. Income from independent labor	
3. Income from capital	
4. Transfers	
5. Transitory income	
6. Unclassified income	
7. Total income	$(1 + 2 + 3 + 4 + 5 + 6)$

Income and Expenditure Data

A total of 500 expenditure items and subtotals are available in the common code. Initial international comparisons will be made at the level of 13 major groups and 34 subgroups. Table 2 describes the expenditure groups and subgroups.

In addition to individual earner's income, income of the unit as a whole is available by breakdown among 6 major functional sources. The sources are listed in Table 2. A distinction between monetary and in-kind income is also available.

International Comparability of the Data

The design criterion for the common international code was maximum comparability of the data, subject to maintaining a reasonable degree of disaggregation. In the face of differing national definitions, some variables lent themselves to transformations that permitted eventual comparison. For example, wide differences in the names of vegetables exist among countries. Many produce items are found only in one or a few countries. Despite these differences, quite good comparability was attained for total expenditure on vegetables across countries. However, differences in definition of some socio-demographic variables were not so readily overcome.

Comparability Through Aggregation

Aggregation was used to attain comparability of income and expenditure categories. An evaluation of the international comparability of the major categories of the income and expenditure data, based on the degree of coverage for the category provided by the items in the country questionnaire, indicates that very good comparability has been attained. Comparability at the next level of disaggregation for expenditures, the 34 subgroups, is also quite good; 90 percent of the subgroups had good to very good comparability.

Noncomparability

Other differences in definition could not be overcome at all. The most serious of these occurred in the socio-demographic data. Definition of the consumption unit varied considerably. Some countries employed the concept of a "secondary unit" to identify a semi-independent consumption unit—for example, married children living with the parents. Some used the "supplementary member" to account for persons who only shared food and/or room expenses. Treatment of domestic employees varied as well; some countries counted them and others did not. Countries distinguishing secondary units and supplementary units would tend to show fewer persons per consumption unit. Those counting domestic employees would show larger consumption units.

Comparisons based on per capita measures of consumption will require care on two accounts. The consumption pattern is susceptible to the type of person included or excluded. Domestic employees, working children, and so on, have expenditures quite different from those of a typical family unit. The inclusion or exclusion of these types of persons and their expenditures and income was systematic, depending on the concept of consumption unit employed. Furthermore,

not all types of expenditure are readily distinguishable. If a person was not counted in the consumption unit, his expenditures and income would not figure in the questionnaire *insofar as it was possible to separate them*. Personal expenditures such as entertainment, clothing, and so forth, were not likely to be counted. Per capita expenditure on living space, durable items, and food, however, would be susceptible to distortion by variations in the count of persons in the unit.

In these instances there was no way to manipulate the information to attain comparability. Means of comparison must be employed which take account of the differences in definition. The differences in definition were somewhat systematic and not so numerous. They could possibly be handled by the use of dummies in multivariate analyses. There was no possibility, however, for *ex ante* transformations to gain comparability as with the income and expenditure data.

DATA PROCESSING

This section describes the data-processing system for the income and expenditure study. The first part sketches some principles applied in the design of the system. The second part describes the preparation of the ECIEL data for international comparisons.

Design Principles for Information Processing

It is useful to consider the design question not from the point of view of information supply but, rather, from the principle of conserving the attention of the researcher and the audience.

Information overload. Empirical research can be envisioned as a combination of data with analysis, given a set of hypotheses. As the quantity of data increases, the human attention required for analysis can become the bottleneck in the process. It is necessary to allocate attention efficiently among the abundant information sources in the data.³

Information overload, a superabundance of data relative to the available attention, aptly characterizes the data-processing context of the ECIEL study. At the research stage, the claim on attention increases with the number of variables. Even greater demands are placed on attention when the data are being handled for the first time. File structuring and data scrutiny precede any analysis. During data preparation, the requirements for attention are proportional to the number of observations as well as to the number of variables. To what extent are the human resources of the project capable of analyzing the voluminous output from a bank of twenty-four thousand observations?

Information condensing. In a situation of information overload, an information-processing system will reduce the net demand on human attention only if it absorbs more information than it produces.⁴ To conserve the scarce resource of attention, the system must be capable of condensing information. The crucial

³ Several views in this section draw on concepts presented by Herbert A. Simon, "Designing Organizations for an Information-Rich World," *Computers, Communications and the Public Interest*, Martin Greenberger, ed. (Baltimore: The Johns Hopkins Press, 1971), pp. 37-63.

⁴ Simon, *op. cit.*, p. 42.

design question is how much and what kind of information it will allow to be withheld from the researcher. The information-processing system should not conserve attention at the expense of the still more salient need for relevance. Performance on this count distinguishes good system design from poor.

During data preparation, the analyst cannot and need not inspect each datum. The system should bring to the attention of the analyst only those data which are obviously erroneous, or those which have a high probability of being in error. Clearly, a tradeoff between accuracy and cost is being made here. The system design can be tailored to shift the balance either way desired.

In response to the decentralization of the fieldwork, the balance was swung toward accuracy in the data-preparation stage. For almost every type of error, the system provided as close scrutiny of the data as possible. Other factors beside the decentralization of the fieldwork reinforced the choice of this option. Any checking that was to be performed had to be done at this stage. The greater the lag between the survey and the checking, the more difficult it would have been to obtain the collaboration of the institutes in checking the original questionnaires.

To facilitate widespread use of the data, we chose to incur all of the fixed costs at this stage and minimize the marginal cost of future applications of the information. This involved the correction of all detectable errors at levels of disaggregation higher than the 47 expenditure subtotals and 6 income subtotals used for the initial international comparison study. Future studies will be feasible at a higher level of disaggregation with virtually no additional fixed cost.

Data Preparation

Figure 1 presents a schematic description of the information-processing system of the study. The units delineate functional steps rather than individual programs. At some points more than one function was performed by a single program. At other points, the intermediate analyses, for example, several programs were necessary to perform the appropriate tests. Not indicated in the diagram are several housekeeping programs used for manipulating the data.

Forms of data. Initially it was expected that the questionnaires could be precoded with cards punched directly from the questionnaire. This did not prove practicable because the questionnaire layout best for field use did not lend itself to keypunching. Consequently the data were passed to coding sheets and some minor conversions and aggregations were performed at this stage. The data were punched locally in all but two cases. They were then forwarded to Brookings for processing.

Stage I of data cleaning. Certain checks of the data were necessary to permit further processing. Cards were passed to tape as soon as possible; this was often done locally. We have had good results with transporting the data on tape and converting the tape for use with the Brookings computer. Disk pack was often used for storage of active data files at Brookings. Tape was used for archive storage of the data at key junctures (data as received, the final clean file, and so on).

Sequence check. Initially card images were checked for proper sequence. An observation in country-specific format occupied from 39 to 60 cards. If the errors in card order were obvious, they were corrected without consultation with the institute. If not obvious, a decision was made either to drop the observation

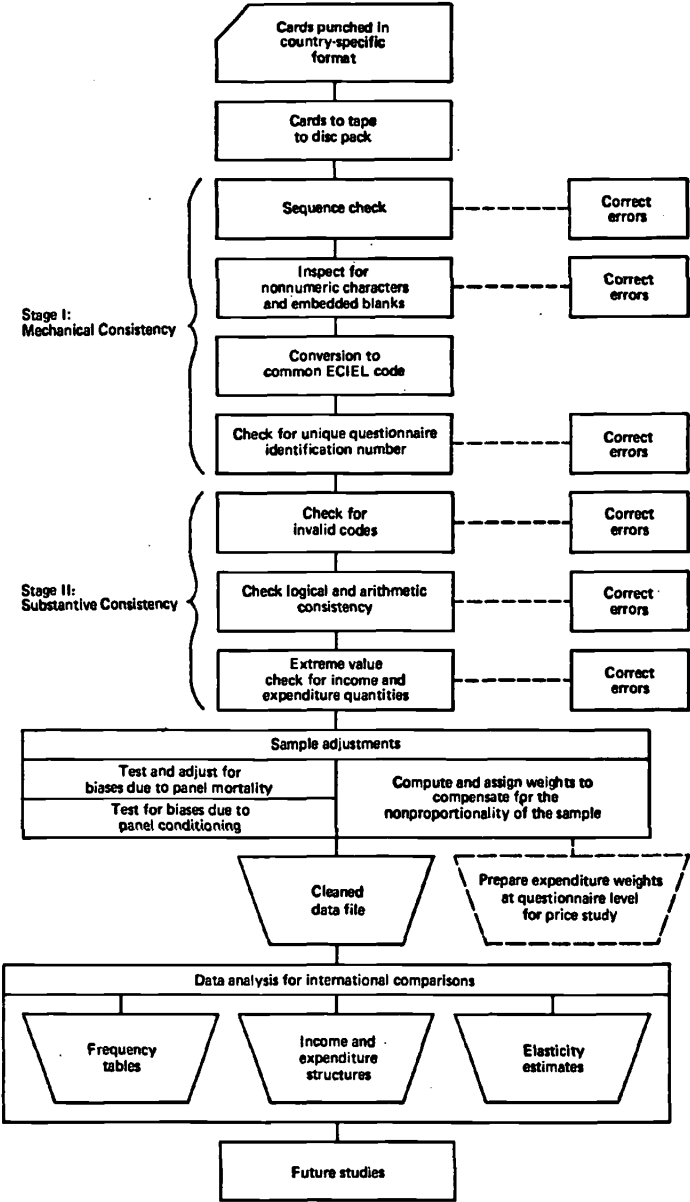


FIGURE 1 INFORMATION PROCESSING SYSTEM FOR ECIEL INCOME AND EXPENDITURE STUDY

permanently, to proceed with the good observations and process the others later, or to delay processing of the entire file until all possible observations were salvaged by checking with the institute. If the work schedule permitted, the last option was preferable. Special update programs replaced the card images in the file.

Nonnumerics and embedded blanks. These errors created different problems. Both were sought at the same time, however, because the programs were based on reading characters in A (alphanumeric) format. The ECIEL code is fully numeric. (Some institutes employed alphanumeric codes for variable formatted observations. See below, *Conversion*.) For faster reading and calculating, I (integer) format was used as early as possible. Nonnumeric characters had to be removed because they would stop execution in this format.

Embedded blanks caused no such mechanical problem. They did, however, give a good indication that a field had possibly been slipped in keypunching. These errors would be found in no other way since, substantively, the blank was handled as a zero. Again, the institute performed double checks on the questionnaires to correct the erroneous data. Virtually all of these kinds of errors were salvageable because no questionnaire and coding sheet should have made it through field supervision checks with alphabets or embedded blanks. Correction at this stage was executed with update programs capable of locating a given variable in a given observation.

Conversion. Original agreements specified that the institutes were responsible for providing the data in the format of the ECIEL common code. However, all of the institutes coded the data in questionnaire-specific form. Some institutes found the conversion beyond their means. In such cases, Brookings prepared the conversion program. Those institutes that did convert the data prepared conversion programs for use as part of the central processing. As discussed above, comparability of the income and expenditure data was obtained through aggregation. In principle, this solution was straightforward. However, several details complicated the conversion.

Variable format.—Most institutes employed a fixed location format for every variable in the country-specific format. This setup was the easiest to handle. However, some institutes used a variable format for the income and expenditure data. This arrangement involved one card for each datum; the format of each card was fixed. An identifying code specified the variable to which the datum corresponded. The number of cards varied with each record, depending on the number of responses obtained. Ostensibly, this format would reduce costs. Few cards are used when there are many zero data. If nonzero responses are expected on only a small number of all the possible data for most observations, this arrangement would economize on cards.

This system enormously complicated the conversion program. Greater conversion costs more than offset any savings on cards and punching. The identifying codes had to be compared with a library to specify the destination of the datum in the common format. The use of an identifying code doubled the opportunity for coding and punching mistakes to cause permanent error. The use of an alphanumeric code by one institute introduced further complication and opportunity for error.

Packed format.—Even in the case of fixed formats, suboptimal format design occurred. All institutes were concerned with reducing card costs. The common

international code and format were being designed at a time when it was still expected that the institutes would be providing the data in common format. Thus, the pressure to reduce card usage influenced the design of the common code as well. As a result, the format was packed. Minimum necessary space was allocated to each variable. Adjacent variables thus had differing field widths. The local pressure for card reduction led to packed formats in the national formats as well. Thus, variable field-width formats were used at two stages. If the Coordination had known that it (rather than the institutes) would convert the data to common code, it would have been possible to employ a more efficient format design for the common code.

Variable field widths caused several difficulties. Indexing of variables was complicated. A fixed number of variables per card would have greatly simplified indexing. Varying the field width increased the opportunity for slipping fields during keypunching. Also, despite the extreme care that went into designing the common format, several field-width overflows occurred.

This complication exemplifies one kind of problem resulting from the organization of the study. Great externalities exist between steps in the data preparation. The manner in which costs were shared between the institutes and the Coordination offered many opportunities for cost shifting to the Coordination. As a result, overall project costs were somewhat higher.

Double counting.—The subtotals used in the questionnaires were not, of course, uniform across countries. All subtotal combinations could not appear in the common code. Insofar as possible, variables were mapped into the common code at maximum disaggregation. Lack of uniformity in the subtotals necessitated great effort to avoid double counting in the mapping.

Nonresponse.—Treatment of nonresponses was especially troublesome. Ideally, this problem should have been resolved in the field; the greatest amount of supplementary information was available at that stage. The time lag between fieldwork and central processing barred the possibility of obtaining any of the missing information detected at the checking stage. Perforce, missing data existed. They had to be given special treatment to permit the use of the existing information. A special code—the field filled with the digit 9—denoted nonresponse. The nonresponses had to be separated out of any computation performed with that variable.

For computing subtotals at all higher levels of aggregation, the missing datum was treated as a zero. (In tabulation and all subsequent computations, observations with missing data were excluded.) The error introduced in the subtotal by this approach depended, of course, on the relative magnitudes of the missing datum and the subtotal. This was not the same across expenditure categories. An individual food item, for example, was probably quite small in relation to the subtotal for food and beverages. An appliance, on the other hand, was likely to represent a significant portion of the subtotal for household equipment. However, the proportion of nonresponses relative to actual zero expenditures here was likely to be much lower than with a food item.

Since the nonresponses occurred at the questionnaire level, while the international comparisons were made at the level of the 47 major groups and subgroups, the nonresponse problem did not complicate the initial computations. Yet to preserve the possibility of working at the most disaggregated level in future

research, the nonresponse codes and the consequent complications of checking for, and separating, the codes had to be borne throughout the data-preparation stage.

Stage II of data cleaning. At this stage, substantive errors were sought and corrected. In all cases, the questionable data were returned to the institute for double-checking. Substantive errors were more difficult to correct, because they may have been recorded as such in the questionnaire. In other words, the only error that could be corrected at this stage, as in Stage I, was a coding or punching error. If the datum was recorded erroneously in the questionnaire it was left as it was or changed to a nonresponse in the data file, depending on its "reasonableness."

Invalid codes and invalid values. Some data could be detected as erroneous by simple inspection. Qualitative data appeared in the file as numeric codes. The most obvious error was the use of a digit not designated for the particular variable. For example, the code for sex of household members was: 1—male, 2—female, and 9—no response. A datum of 0, 3, 4, 5, 6, 7, or 8 on this variable was invalid. A coded variable that employed all the digits could not, of course, be checked for this type of error.

Quantitative data covered a much wider range. Any number except a code (field filled with 9's—no response; field filled with 9's except 8 in the right-hand column—not asked in questionnaire) was a valid value. Expenditure and most income variables were nonnegative quantities. All variables that should have had positive responses only were checked to insure that no negative data appeared.⁵

Logical and arithmetic consistency. The interrelatedness between certain variables was exploited to check for erroneous data. If the data for the variables were not consistent with the a priori relationship, at least one of the data had to be erroneous.

A multitude of relationships existed among the socio-demographic variables. All possible logical relationships, approximately 50, were checked. Some typical relationships are presented as examples:

- (1) children must be younger than the parents by at least 13 years;
- (2) educational level must be consistent with age, i.e., a child cannot be recorded as having a university education;
- (3) the number of persons must add up to the family size reported;
- (4) both rent and mortgage payments cannot be reported for the same dwelling;
- (5) expenditures can be made only for utility services (gas, electricity, and so on) installed in the dwelling.

Furthermore, all of the income and expenditure data were subjected to an adding-up check. Subtotals had to be greater than, or equal to, the addends. The greater-than inequality allowed for the possibility that the breakdown for certain expenditures was not known but the category total was known and reported. In

⁵ This test was performed with a univariate distribution program. To test for negative values, the valid limits were established at zero and the maximum for the particular field width. The program output identified all observations having negative values. The program also divided the range between zero and the maximum valid value into twenty equal intervals and displayed the resulting distribution. The distributions proved to be helpful aids in interpreting the extreme value test (see below).

this case, the reported total exceeded the sum of the reported addends. In the consistency checks, nonresponses were handled as zero.

Although it would have been conceivable to check for consistency over time for those consumption units reappearing in the panel, such checking was not carried out. Problems in identifying family members (they might not be listed in the same order in each interview), possible changes in employment status, and defining "consistent" income or spending over time promised to make the check more costly than the benefits derived.

Extreme value test. As far as we are aware, this was the first time that regression methods were used in a checking procedure. This approach constitutes a considerable increase in the scrutiny with which survey data are reviewed on a large scale. Previously, the most detailed checks established valid ranges for expenditure quantities normalized by total expenditure. For example, all observations whose percentage expenditure on food did not lie between, say, 30 percent and 60 percent would be reviewed in detail. Other determinants of expenditure level were not taken into account.

The method employed here allowed for multiple explanatory variables for expenditure level of each of the 34 expenditure subgroups and 6 income subgroups. The test used a simple arithmetic regression model. The explanatory variables for each income and expenditure category were chosen on the basis of a priori notions of the determinants of spending and income. The advantage in employing multiple explanatory variables was the reduction in the number of observations brought into question. For example, if a household showed a high percentage expenditure on food, but the number of members was high, the high percentage was likely justifiable. With the regression test, this observation would not have been singled out for inspection. We desired a method that could rapidly scan the data file and identify "unreasonable" incomes or expenditure for further inspection.

Unreasonable in this case amounted to unreasonably high. We were concerned about errors which will bias the statistical estimates. They could be biased positively or negatively by observations with erroneously overreported or underreported expenditures and incomes. However, the danger from errors on the low side was less severe. Since income and expenditure items were positive, all negative quantities had been removed by checking. Errors were thus bounded on the low side (at zero) but not on the high side.

Up to this point in the data preparation, all erroneous data that could possibly have been detected were identified. This test, however, depended on the likelihood of a datum being erroneous. Only further checking with the questionnaire could confirm that it was a coding or punching error. There was latitude for a choice between certainty and cost. Absolute certainty on coding or punching errors would have required the inspection of every datum. Because we were seeking only extreme errors, we were able to focus attention on a relatively small number of observations.

The method compared the actual expenditure or income value with that predicted by the regression equation. It then identified all observations whose residual (actual value minus predicted value) lay more than three standard deviations away from the regression surface. Three standard deviations was an arbitrary

choice. It was chosen to reduce the number of observations demanding human attention while picking out those erroneous values large enough to seriously bias parameter estimates.⁶

Only those observations having nonzero expenditures and incomes for the category in question entered the equation. We had no way of knowing whether they were correctly recorded as zero. We could not impose the condition that because a household had certain characteristics it *must* have made the expenditure or received the income. The presence of observations with zero for the dependent variable would have served no useful purpose and would have tended to confuse the interpretation of the regression equation.

The procedure gave very good results the first time it was employed on a data file of 2,949 observations from one country. In checking the 34 and 6 subgroups, 932 "extreme values" were identified. The percentage of observations identified per equation varied between 0.5 percent (food and beverages) and 2.6 percent (medical care). After double-checking on the questionnaires, 16 percent of the 932 extreme values were found to be errors introduced by coding or punching. The question remained of what to do with the "extreme values" that were not punching errors.⁷ The easiest procedure was also the safest. Doubtful values were left as they were, unless they were extreme enough to bias the data.⁸ Even if we were dealing with all "true" data, depending on the shape of the distribution of the residual, the 3 standard deviation test could have identified up to 11 percent of the observations as "extreme." In the first actual use of the test, none of the 84 percent of the observations inspected but not having punching errors were extreme enough to warrant replacement; they were all left as they were in the data file.

Few extremely low values were identified by the test. This was likely due to the fact that income and expenditure distributions are typically skewed to the right. Hence, more observations fell within 3 standard deviations on the low side than on the high side. If the variance of the residuals was high, the 3σ limit could fall below zero and no low values would have been detected (negative values had already been purged).

Sample adjustments. Up to this point, efforts were directed to correcting individual observations. Three possible sources of bias due to the sampling survey remained to be adjusted for: nonproportionality of the sample, panel mortality,

⁶ If all the data were correct, up to 11 percent of the residuals could still have fallen outside 3 standard deviations. If in addition, the distribution of the residuals was approximately normal, as few as 0.3 percent of the observations may have fallen outside. Since the distribution of the residuals was likely to be unimodal with high contact, about 1 percent was a good guess for the proportion of valid observations lying outside the band. Working with a file of three-thousand observations, this would single out about 30 observations per equation for detailed inspection. From the point of view of information condensing, this is about all one would need to spot very serious outliers.

It is unlikely that as many as 30 observations would be an order of magnitude larger than the mean without our already having known about it. Univariate distributions of each subtotal were available at this stage. They were used in conjunction with the extreme-value test. If a large number of observations was noted at the high end of the distribution, it would have been easy to set the limit at 2 standard deviations for that particular variable and thereby scrutinize more observations with large values.

⁷ Originally we speculated on the desirability of replacing the datum by the predicted value plus a random component. This procedure was rejected because it would have amounted to homogenizing some of the data (albeit to a small degree—less than 3 percent of the data) according to our a priori and oversimplified notions of the determinants of expenditure.

⁸ "Extreme enough" would seem to be an observed value an order of magnitude larger than the mean of the dependent variable.

and panel conditioning. To date, only the nonproportionality adjustment has been carried out; the test and adjustments for the panel have not been completed. The sample adjustments are not particular to the ECIEL study; they are normally made on any survey data, regardless of the degree of decentralization of the fieldwork.

Nonproportionality. All sample data in this study had to be corrected for nonproportionality. The samples were designed to be stratified by income level with different sampling fractions employed. The most common method for adjustment involved the use of exogenous information on the distribution of income of the population. The income distribution resulting from the sample was used to determine appropriate weights to scale the sample data up to the population income distribution. These weights were assigned to the individual observations as a datum. Whenever a weighted mean was required, the computation programs brought in this datum as the weighting factor.

Panel mortality. As indicated in Table 1, eight of the eleven countries employed a panel. The panel is a subsample of families that was interviewed in each of the four intervals. Basically, the panel is used as a control group for variations in consumption through the year. If differential (by socio-demographic characteristics of the household) mortality occurred, the panel may not provide a good control basis. As a test, other subsamples were designated for once- and twice-interviewed families in different intervals. A series of χ^2 tests are performed to test for differences in composition of the subsamples and differential mortality in the sample over time. The cross-tabulation program written for ECIEL provides the χ^2 statistic for each table. Control variables checked here are size of household, employment status of head, age of head, and total income. If significant differences in composition of the panel subsamples across intervals are detected, differential weighting is applied to the observations to compensate.

Panel conditioning. The periodic reinterviews may make panel households more aware of their income and expenditures than they otherwise might be. As a result, they might have altered their spending or reporting practices to differ from those of comparable households being interviewed for the first time. The tendency for income and expenditure magnitudes to rise due to inflation further distorts the sample results over time. To test for panel conditioning, an analysis of variance is performed to determine significant variations in mean expenditure. The objective of the analysis of variance is to test for the significance of observed interaction between subsample and the control variable, and for the significance of the differences in mean expenditure among subsamples.

The control variables are: subsample number, quarter, and number of interview (first time, second time, and so on). Mean expenditures are to be tested for all 13 major expenditure categories (see above, Table 2). Three income variables—total income, wages, and income from capital—are also tested. The tests for panel mortality and conditioning have not yet been completed. Unlike panel mortality, conditioning effects are not readily adjusted for by weighting. The precise nature of the adjustment procedure for panel conditioning will be determined after initial tests are carried out.

Summary of data preparation. Throughout the data-preparation stage, it was necessary to balance accuracy against cost. For most checks, this amounted to

performing, or not performing, the check. Such checks, by their nature, detected *only*, but *not all*, errors. It is evident from the descriptions of the checks that all errors could not be detected. The process only detected erroneous *and* inconsistent data. Data that were erroneous but consistent with the other information in the observation slipped through the tests. If a certain check was desired, it would focus attention on *only* errors, and all errors that were *detectable*. If such a check were deemed too costly for its contribution to accuracy, then *no* error of that type would be detected.

The extreme-value check differed in this respect; some of the extreme values may not have been erroneous. Within the boundaries of the test, it was possible to shift the balance toward inspection of more dubious information or toward economizing on human attention.

For data preparation as a whole, the balance was decided in favor of maximizing accuracy. This decision was largely made in response to the decentralization of the fieldwork. The desire to incur now as much of the fixed cost of data preparation as possible and to reduce the marginal cost of future studies complemented the decision to perform as many checks as possible.

The first two checks—sequence and nonnumerics—would have had to be performed in any study. They involved mechanical errors that would have involved nonsense processing in the first place, and interruption of program execution in the second. The check for embedded blanks was a good way to locate slipped fields in coding and punching. Its technical interdependency with the nonnumerics check permitted its incorporation at a relatively low marginal cost.

Determination of the point at which conversion was carried out involved subtle tradeoffs. To maximize the opportunity for uniform treatment of the data, conversion should be made as early as possible. On the other hand, certain kinds of errors—cards out of sequence and nonnumeric characters—had to be removed to permit conversion. Also, the earlier the checking was carried out, the closer to the source the error could be located, and backtracking through the intervening transformations was minimized.

Because of its technical interdependency with the nonnumeric check, the embedded-blank check was performed before the conversion. Since the socio-demographic data generally needed no transformation—their conversion mainly consisted of reformatting—the location of the logical consistency check was not critical from the standpoint of backtracking. Conversion was located before the logical and arithmetic consistency checks so that these could constitute a double check on the conversion as well as a check on the data themselves. That is, the logical and arithmetic checks tested the reformatting and the adding-up done by the conversion, in addition to the consistency of the data. This benefit compensated the backtracking necessary to locate errors in income or expenditure data at the questionnaire level. Since the extreme-value test was run on income and expenditure subtotals, it perforce had to be located after the conversion, even though corrections had to be sought at the questionnaire level.

One feasible check was not incorporated in the data-preparation system. A logical-consistency and extreme-value test of sorts was conceivable for the panel families interviewed more than once. These checks would have taken the form of imposing consistent behavior of a given household over time. The difficulty of

defining consistency over time, and the low marginal benefits of the checks, resulted in their omission.

Invalid codes and invalid values did not have to be removed from a mechanical standpoint. Tabulations could have classified them as "other." However, they could have been numerous and a reduction in the effective sample size would have resulted. Furthermore, negative expenditure values could have seriously biased the statistical estimates.

Logical and arithmetic consistency were not essential from a mechanical point of view either. Again, however, such inconsistencies could have been numerous and could have adversely affected the results. Also, as noted above, these checks served as an effective double check on the conversion.

The extreme-value test was incorporated to prevent severe biases that would result from the presence of extremely large values not counterbalanced on the low side, due to the lower bound of zero. Here, however, it was possible to employ the principle of information condensing and reduce the number of observations requiring scrutiny. Only those values capable of severely affecting the results, and with a high probability of error, were inspected.

The nonproportionality adjustment was essential. The tests and adjustments for panel mortality and panel conditioning constitute a refinement attempting to adjust for effects in the interview process itself. This contrasts with the other tests in that they essentially discover errors introduced in the transmission of the information after the interview stage.

A review of the benefits of the extensive cleaning undertaken as part of the data preparation completes the summary. Even if all errors could not be detected by the data-cleaning system, the results of the cleaning probably give a good indication of the overall quality of the field response and coding operation. The number of undetectable errors in the data is probably highly correlated with the number of inconsistencies identified by the checking. Inconsistency rates can be employed as a rough indicator of the comparative quality of the data.

Prior cleaning and inspection will enhance the validity of the results of the comparative studies. The data will remain useful for a long time after the ECIEL international comparisons are carried out. Hopefully, the data will come to be used by a great many scholars for a wide variety of national and international studies. The data will have greater usefulness if accompanied by indications of their quality. Over the long run, the benefit from careful data preparation should exceed its costs.

CONCLUSION

On the basis of our experience, we can state a few technical criteria for micro-analytic survey work that should hold under most conditions. No attempt should be made to economize on the use of cards. This is the number one false economy for data processing. Data packing and coded free formats enormously complicate the later processing.

Alphabetic codes should be avoided. They complicate the input formatting and then require cleaning checks to insure that they do not appear in fields where they do not belong.

Extra effort should be expended to solve nonresponse problems in the field. This measure is closely related to performing more consistency checks at the field level. In a joint study, the time lag between fieldwork and processing is great enough to preclude the possibility of correcting errors by return visit. The return visit is the optimal way to correct nonresponse and inconsistent information. The use of later automated checking does not obviate the need for good field checking procedures.

The benefits to permitting slight differences in questionnaire design are less than the costs of the additional complication. Difficulties due to differences in household terminology can be overcome with clarifying notes in the questionnaire. Often the solution is simpler; if a given item has a different name in another country, it can appear in the questionnaire under its local name. But it should go in the *same* location with the *same* code number as in other countries. The problem of nonexistence of certain items in household budgets or nonavailability is likewise tractable in a common questionnaire. Space for such items can be allowed in the format and the item blacked out in the questionnaire. The extent of differences in definition and nonavailability was overstated by the participants. The extra questionnaire space and cards due to a uniform questionnaire constitute the most expendable resource in a comparative study. These general criteria, however, are not sufficient guides for effective large scale microanalysis in either a national or comparative context.

The major consequence of these unforeseen factors was additional cost. Few coordination problems resulted in permanent impairment of the data. The data-preparation procedures have succeeded in overcoming most of the incompatibilities. The comprehensive coverage of the LAFTA region more than compensates for the additional costs necessary to obtain the participation of all countries. Indeed, the body of microanalytic data proceeding from the study promises to be the most compatible, comprehensive, and reliable yet assembled for Latin America.

Better information on costs could be generated in studies currently under way. We have attempted to identify and clarify the numerous aspects of large-scale information processing about which little information is currently available. We state a few open questions: How should the work be staged for optimum balance between machine and human resources? In how large increments should results be sought? (What portion of intermediate results are ultimately discarded?) What are the payoffs to data cleaning? (A reverse validation on uncleaned data would indicate how much the results differ.)

We would urge that additional resources be made available to individual research projects for the express purpose of investigating concurrently the computing aspects of the research. Such seed resources could draw out a body of organized methodological guidelines that would serve to improve the efficiency of large-scale economic-research computing.

APPENDIX

Institutes Participating in the ECIEL Household Income and Expenditure Study
Argentina: Centro de Investigaciones Económicas, Instituto Torcuato Di Tella, Buenos Aires.

- Bolivia:** Instituto de Investigaciones Económicas, Universidad Mayor de San Andres, La Paz.
Instituto de Estudios Sociales y Económicas, Facultad de Ciencias Económicas, Universidad Mayor de San Simón, Cochabamba.
- Brazil:** Instituto Brasileiro de Economía, Fundação Getulio Vargas, Río de Janeiro.
- Chile:** Instituto de Economía y Planificación, Universidad de Chile, Santiago.
- Colombia:** Centro de Estudios de Desarrollo, Universidad de Los Andes, Bogotá.
- Ecuador:** Instituto Nacional de Estadística, Quito.
- Mexico:** Centro de Estudios Económicos y Demográficos, El Colegio de México, México, D.F.
- Paraguay:** Centro Paraguayo de Estudios de Desarrollo Económico y Social, Asunción.
- Peru:** Centro de Investigaciones Sociales, Económicas, Políticas y Antropológicas, Universidad Católica del Perú, Lima.
- Uruguay:** Instituto de Estadística, Facultad de Ciencias Económicas, Universidad de la República, Montevideo.
- Venezuela:** Centro de Desarrollo, Universidad Central de Venezuela, Caracas.
Departamento de Estadísticas, Banco Central de Venezuela, Caracas.

